Just Say “No” to the Johnson Transformation

One significant weakness in data analysis is use of the Johnson Transformation. While never using it myself, there have been attempts by others I have witnessed. They end up stuck and asking for our assistance. Each time, the Johnson Transformation application was not appropriate. Why is it compelling for people to use? What could be done instead? We will answer these questions for you.

What is it?

The Johnson Transformation is actually a “family” transformation which has 4 components. The first component is no transformation, or just analysis of the raw data. Clearly, there could be no issue with this approach as long as the proper tool is used and the data is normally distributed.

The second component is a natural log transformation. This is acceptable for natural processes and occurs in scientific and engineering data. We need to understand the data skew and have enough data to define it well. Again, no problem here if applied appropriately.

A third component of the Johnson Transformation is the “logit” transformation. This is for data with firm physical boundaries. There needs to be solid support for the boundaries chosen and the boundaries need to be an input from a knowledgeable user. In some data analysis packages, this is just a mathematical game which can correct the data to normal. However without real support for the boundaries chosen, this is not a solution that can provide a predictive model or equation. Which means it is not good for decision making. We have started to get in trouble with application of the Johnson Transformation with this third component.

The final component of the “family” is a hyperbolic arcsin transformation. This is a mathematical operation which almost always seems to make data appear normal. What is the basis of it? It appears to be just a mathematical trick. The coefficients involved are calculated in a manner to fix a numbers problem. However it is not a good scientific model building approach. It does not improve our technical understanding or help us make better decisions. In one case studied, the raw data was actually bimodal (two distinct peaks). Breaking the data down with a specific contributing variable, the data was now normal. This had supportable, scientific reasons. Thus, we could better understand the raw data and could make more confident decisions. The Johnson transformation would have limited our ability to learn about our system.

Why is it compelling to use?

Very simply, using the Johnson Transformation is compelling because mathematically it seems to almost always make data appear normal. This allows us to check the “data is normal” box and to use the conventional statistical tools for normal data. Because of this, it is quick and easy. Digging deeper or understanding alternatives takes more work. People would prefer to avoid excessive work (which is natural) but especially when they are overwhelmed with other activities. The negative impacts of such a choice are not understood, so people proceed to blindly use this tool.

Why again is it so bad?

The trouble is the Johnson Transformation prevents our scientific understanding which limits our ability to make strong decisions. The data is what it is, so do not modify it and hide the reality of what is happening. Better understanding allows us to understand the science behind our system and to anticipate scaling rules to apply the learning to other applications. Learning is good, but leveraging our learning accelerates our overall knowledge acquisition even further. This accelerates not just the current project, but also other related projects. As they say, knowledge is power.

What should we do instead?

Instead of using this family transformation, we could use the natural log transformation if and when it makes sense. We could use the individual transformation for logit, with intelligently selected boundary values. Never use mathematically calculated boundary values for logit as discussed earlier.

However, we should never use the “hyperbolic arcsin” option in scientific model building. Digging deeper and finding an appropriate stratification of the data improves our learning and enables us to make better decisions.

Of course, other transformations could also be used beyond what we have already discussed. Many other transforms exist including square root and “power” versions. Further modeling (inclusion of other input variables) could also be done to find an appropriate match. Do not be overwhelmed with the variety of options as time with your data will narrow your options and obtain a consistent approach.

Conclusion

Data analysis and model building can be a complicated activity. Some over-simplify the situation to make things easy on themselves. They will do this by assuming normal data and using the simple and standard tools. Others misuse more complicated tools such as the Johnson Transformation. A lack of understanding of the tools leads to not understanding the technology being investigated. Just because it works does not mean it was selected appropriately.

Do not risk your project and your business by taking the easy way out and using the wrong tools at the wrong time. Using an expert in New Product Development and data analysis can help make better decisions and ensure down-stream surprises are limited. This can also provide a positive impact on other projects. Leveraging our new product knowledge provides benefits far into the future.

- - Perry Parendo

Contact Information:

Perry Parendo
651-230-3861
Perry@PerrysSolutions.com
www.PerrysSolutions.com